Research Journal of Applied Sciences, Engineering and Technology 8(3): 341-345, 2014
DOI:10.19026/rjaset.8.978
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: February 11, 2014 Accepted: June 08, 2014 Published: July 15, 2014

Research Article

HADOOP+Big Data: Analytics Using Series Queue with Blocking Model

S. Koteeswaran, P. Visu, K. Silambarasan and R. Vimal Karthick
Department of CSE, Vel Tech RR and SR Technical University, Chennai-600 062, India

Abstract: Big data deals with large volumes of tons and tons of data. Since managing this much amount of data is not in the mere way for the traditional data mining techniques. Technology is in the world of pervasive environment i.e., technology follows up with its tremendous growth. Hence coordinating these amount of data in a linear way is mere little difficult, hence we proposed a new scheme in order to draw the data and data transformation in large data base. We extended our work in HADOOP (one of the big data managing tool). Our model is fully based on aggregation of data and data modelling. Our proposed model leads to high end data transformation for big data processing. We achieved our analytical result by applying our model with 2 HADOOP clusters, 4 nodes and with 25 jobs in MR functionality.

Keywords: Big data, blocking queue, data rendering, hadoop, map reduce

INTRODUCTION

In real world, most of the online stores, social forums, e-commerce, online stores etc., stores the enormous amount of data represented in terms of on-click user generated terms, To activate the huge resource utilization in terms of big data the user defined inputs such as text, images, videos etc., are consider in term (Sagiroglu and Sinanc, 2013; Ji *et al*., 2013; Lee *et al*., 2013; Farhana *et al*., 2013). With rapid development in networking technologies the distributed client server architecture is used typically in order to increase the performance vectors. For these evaluation for big data's various tools are used and various vendors which utilizes these tools for processing the user inputs, some of the tools are mahout, Hadoop, cloudera etc., (Kezunovic *et al*., 2013). Here we use HADOOP tool for processing the data using MR methodologies. Map reduce function are used in various cluster for parallel processing of data and data sets.

What is big data analytics? Simple processing of these huge amounts of data is not easy to process and pre-process, since the data's are to be stored and retrieved quickly (Jensen, 2013). Traditional mining methodologies are out dated nowadays due to its lack in storage option, additional hardware utilization and increased level of processing time etc. Hence here comes the actual picture of big data. Big data supports remote cloud, online processing and it has enormous features. Now a day's lot of mining vendors are switched to big data analytics.

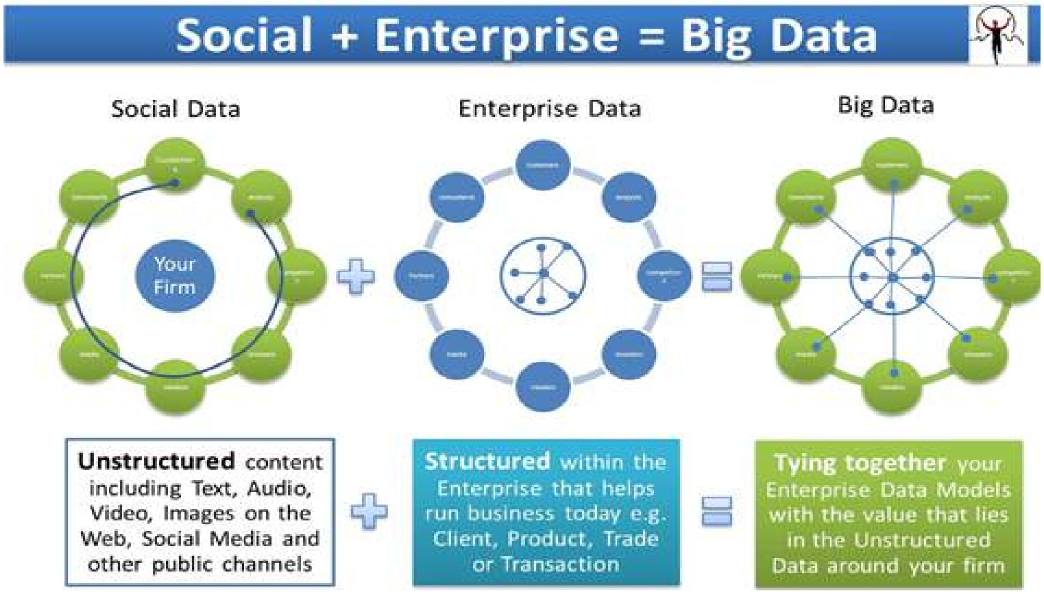
How data are transformed into big data's, this simple question was demonstrated clearly in Fig. 1. It denotes the vital utilization of social media data and enterprise data is transformed to big data (WATALON, www.watalon.com.). It supports all the Business roles in term of Lehmann statement with accuracy and zero defects with large analytics in order to achieve the actionable knowledge and insights. Since the data stored and processed statistically is subjected to various issues, statistical computing is in emerging wide because of these big data and its sophisticated knowledge analytics (Das and Ranganath, 2013; Jensen, 2013; Koteeswaran *et al*., 2012; Koteeswaran and Kannan, 2013; Vera-Baquero *et al*., 2013).

These enterprise models focus is towards the structuring the statistical data in linear way, hence it is to be acquired by the knowledge resource and hence data transformation is in need. Here comes the picture of "data transformation" to lead these kinds of issues. Typical architecture to evaluate our proposed model is demonstrated clearly in the Fig. 2.

MATERIALS AND METHODS

Hybrid model: To overcome the challenges like linear transformation of data and data translation. We proposed a new hybrid scheme, in this scheme data is transformed and reduced into linear maps by means of map reducer. Here various map reducers are deployed to predict the linear way of data; hence data preserved is translated/transformed and processed by HADOOP. Each block holds the stack of each rack.

Corresponding Author: S. Koteeswaran, Department of CSE, Vel Tech RR and SR Technical University, Chennai-62, Tamil Nadu, India, Tel.: +91 9884378785



Unstructured content including Text, Audio, Video, Images on the Web, Social Media and other public channels

Structured within the Enterprise that helps run business today e.g. Client, Product, Trade or Transaction

Tying together your Enterprise Data Models with the value that lies in the Unstructured Data around your firm

Fig. 1: Big data in terms of enterprise edition ((Sagiroglu and Sinanc, 2013), (WATALON, www.watalon.com.))

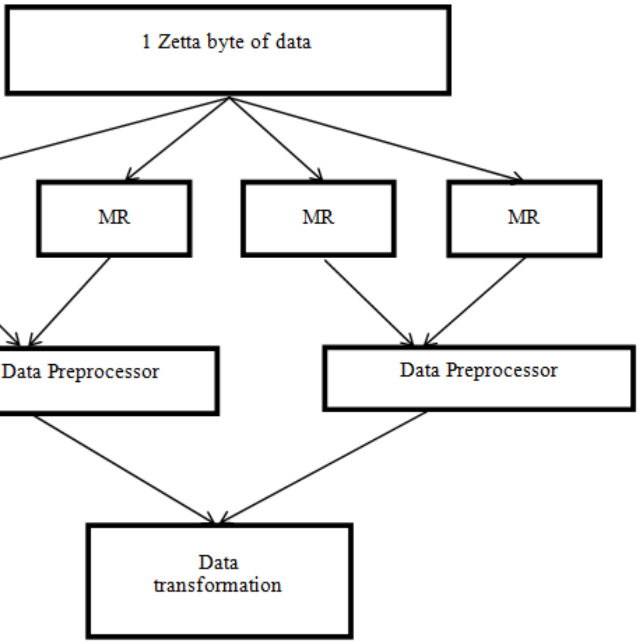


Fig. 2: Architecture data rendering

Data nodes in HADOOP clusters: Linear transformation of data and processing those non-linear segments is one of the vital challenge in security scrutinizes, since these transformation leads to replication of data in blocks, racks etc., these fields leads to decrease in performance, reduced stats in execution during retrieving, virtualization will not adopt to non-linear data sets. Clustering these nodes and creation of nodes will increase the performance and used for synchronizing huge amount various data's into variant data models.

Map reduce functional model: Map reduce functional model is used to process various data models in terms of paralleling and distributing it over huge amount clusters. Map reduces focus on various roles such as data filtering and sorting. Here our approach is to process the data parallels and to distribute it over large Hadoop clusters.

How parallelizing comes into the picture (MR): In DFS, each data blocks are key node identity, in which the nodes particular identity as well node integrity is

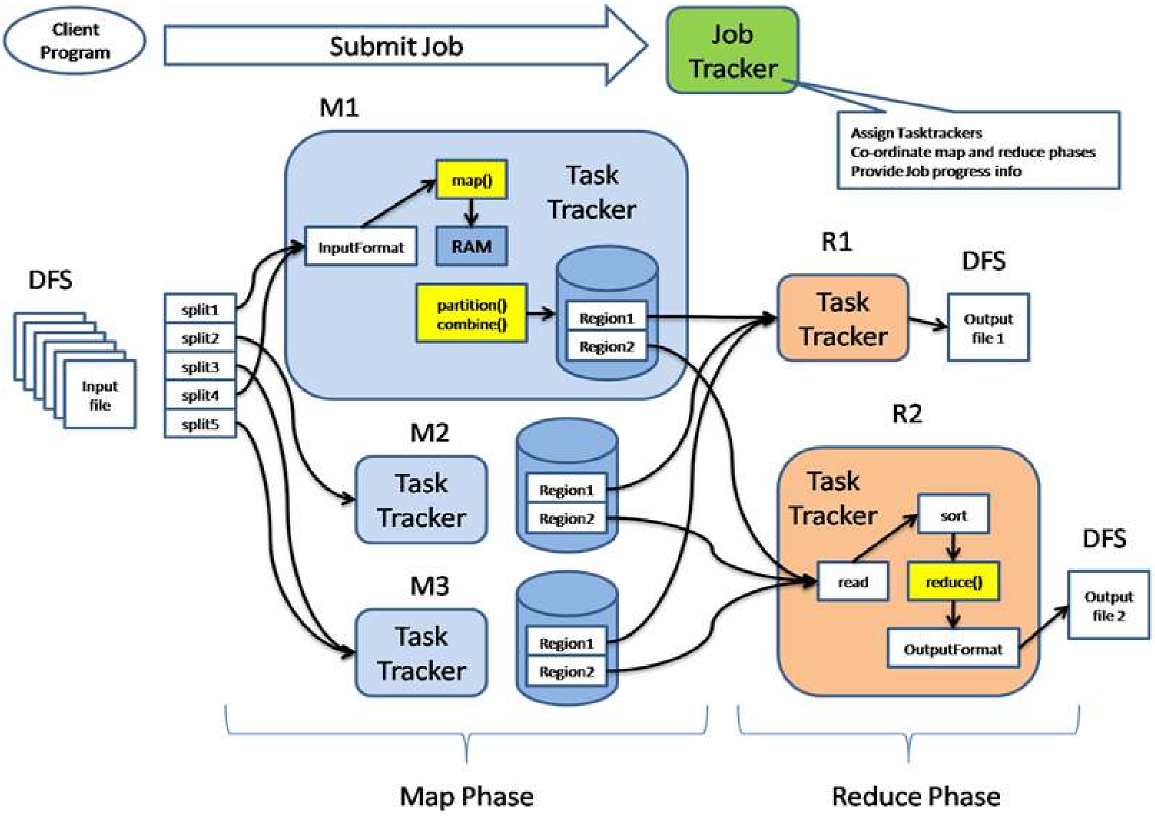
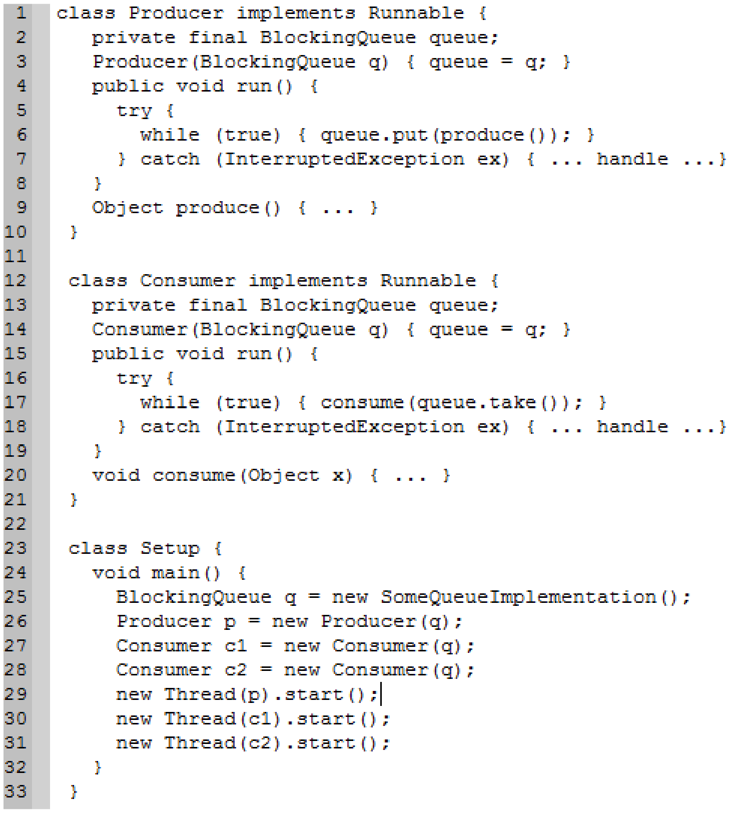


Fig. 3: Map reduce model for performing parallel data distribution in Hadoop clusters



```
class Producer implements Runnable {
  private final BlockingQueue queue;
  Producer(BlockingQueue q) { queue = q; }
  public void run() {
    try {
      while (true) { queue.put(produce()); }
    } catch (InterruptedException ex) { ... handle ...}
  }
  Object produce() { ... }
}

class Consumer implements Runnable {
  private final BlockingQueue queue;
  Consumer(BlockingQueue q) { queue = q; }
  public void run() {
    try {
      while (true) { consume(queue.take()); }
    } catch (InterruptedException ex) { ... handle ...}
  }
  void consume(Object x) { ... }
}

class Setup {
  void main() {
    BlockingQueue q = new SomeQueueImplementation();
    Producer p = new Producer(q);
    Consumer c1 = new Consumer(q);
    Consumer c2 = new Consumer(q);
    new Thread(p).start();
    new Thread(c1).start();
    new Thread(c2).start();
  }
}
```

Fig. 4: Sample code for series blocking queue

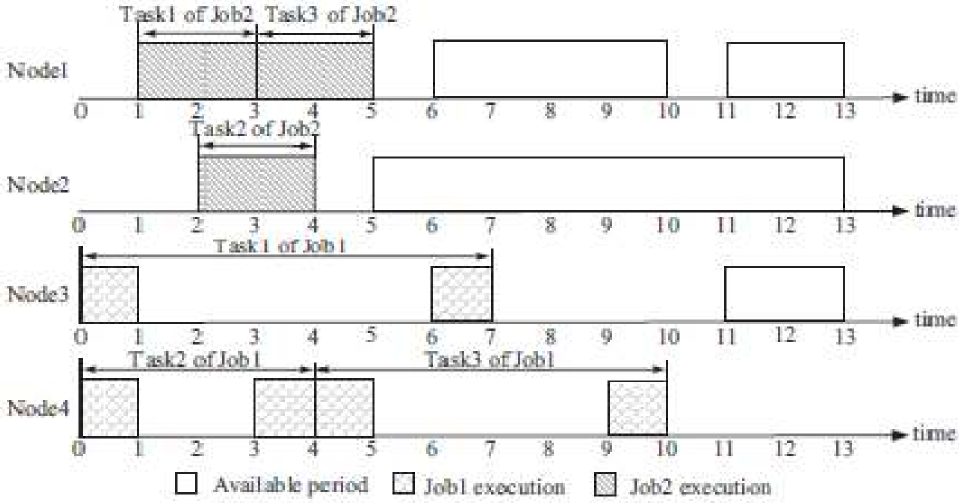


Fig. 5: Parallelizing the jobs in MR in HDFS

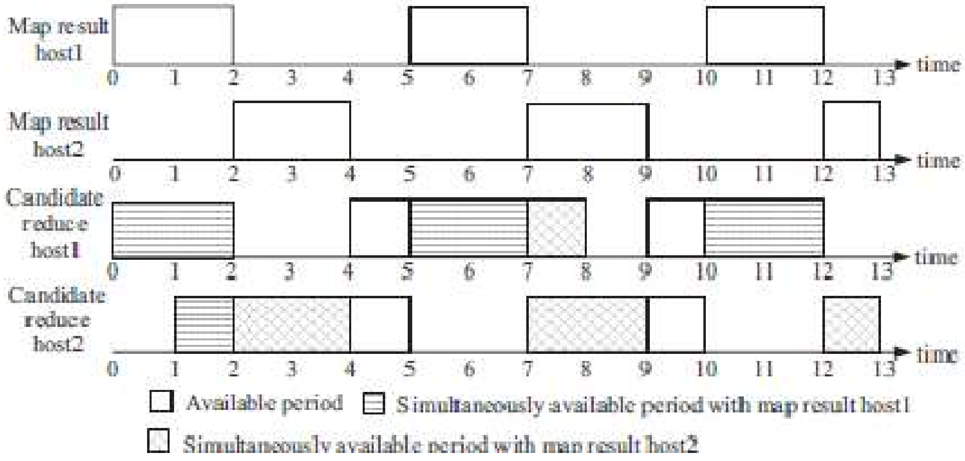


Fig. 6: Available resource between max (hosts) and MR

checked by the HDFS, since the name node performs the fault tolerance problem in terms of name conflicts, each record blocks are investigated by the same. In order to parallelize the whole procedure, we used HDFS based generic algorithm which was inbuilt and used by the Hadoop ecosystem, here we achieved our parallelizing by reversing the series blocking queue in order to distribute the whole data sets in various clusters (Fig. 3).

Algorithm for parallel distribution of data's in hadoop nodes (clusters):

- Create Hadoop clusters
- Assign node identity and integrity
- Process the input (Zeta byte of data's)
- Execute the used defined MR code
- Shuffle and sort the pre-processor code of MR
- Execute the user define reducer code
- Achieve the result with MR functionality
- Distribute the data's in the Hadoop clusters

Blocking queue in our model: Reversing the blocking queue model is achieved in order to increase the performance and fault tolerance during data transformation (Fig. 4).

RESULT ANALYSIS

What we have performed for testing our data sets in order to achieve tons of data in singleton result, we have used the cloudera data set for testing and validating our code and achieved with 92% in data transformation. Here the Fig. 5 and 6 denotes the resource utilization of MR and parallel task in job tracker in Hadoop pre-processor. In Fig. 5 the node 1, node 2, node 3 and node 4* are in parallelizing event and the incoming job in task scheduler and job tracker will be of in the pattern of queue, here FIFO method is used to handle all the jobs, if the priority jobs request the Hadoop node first then the job in execution is pre-empted and rescheduled in the job tracker where the task tracker holds of the same for a while and then the job is released from the queue. The job listed in resource utilization should be based on CPU resource utilization time.

CONCLUSION

In this study we proposed a new model for data transformation using series queue blocking, it is our reputation to find the large data sets are processed and distributed among clusters, Modifying the MR functionalities in kernel level for parallelizing the data distribution is major form factor will reduces fault and increase the performance of the clusters. An analysis of the huge data sets are made and results demonstrating the field of business values have been developed and demonstrated the show case of the novelty in this approach. With the interest in rapidly developing form factors, our next work is towards the integrity level in the big data nodes. In future we would like to enhance our work in node identity and integrity in order to increase the security level in clusters and in Hadoop framework.

REFERENCES

Das, A. and H.S. Ranganath, 2013. Effective interpretation of bucket testing results through big data analytics. Proceeding of IEEE International Congress on Big Data (BigData Congress), pp: 439-440.

Farhana, Z., M. Patrick, Z. Ying, B. Michael, G.S. Femida and A. Ashraf, 2013. Towards cloud-based analytics-as-a-service (CLAaaS) for big data analytics in the cloud. Proceeding of IEEE International Congress on Big Data (BigData Congress), pp: 62-69.

Jensen, M., 2013. Challenges of privacy protection in big data analytics. Proceeding of IEEE International Congress on Big Data (BigData Congress), pp: 235-238.

Ji, Y., L. Tong, T. He, J. Tan, K.W. Lee and L. Zhang, 2013. Improving multi-job mapreduce scheduling in an opportunistic environment. Proceeding of IEEE 6th International Conference on Cloud Computing. Santa Clara, CA, USA, pp: 9-16.

Kezunovic, M., L. Xie and S. Grijalva, 2013. The role of big data in improving power system operation and protection. Proceeding of IREP Symposium-Bulk Power System Dynamics and Control-IX Optimization, Security and Control of the Emerging Power Grid (IREP, 2013). Rethymnon, Greece, pp: 1-9.

Koteeswaran, S. and E. Kannan, 2013. Analysis of Bilateral Intelligence (ABI) for textual pattern learning. Inform. Technol. J., 12(4): 867-870.

Koteeswaran, S., J. Janet and E. Kannan, 2012. Significant term list based metadata conceptual mining model for effective text clustering. J. Comput. Sci., 8(10): 1660-1666.

Lee, C., C. Chen, X. Yang and B. Zoebir, 2013. A workflow framework for big data analytics: Event recognition in a building. Proceeding of IEEE 9th World Congress on Services (SERVICES). Santa Clara, CA, USA, pp: 21-28.

Sagiroglu, S. and D. Sinanc, 2013. Big data: A review. Proceeding of IEEE International Conference on Collaboration Technologies and Systems (CTS), pp: 42-47.

Vera-Baquero, A., R. Colomo-Palacios and O. Molloy, 2013. Business process analytics using a big data approach. IT Professional, 15(6): 29-35.